Indexierungstheorie für Linguisten

Zu einigen natürlichsprachlichen Zügen in künstlichen Indexsprachen

Engerer, Volkmar Paul

Published in:
Dialekte, Konzepte, Kontakte

Publication date:
2014

Document version
Tidlig version også kaldet pre-print

Citation for published version (APA):
Engerer, V. P. (2014). Indexierungstheorie für Linguisten: Zu einigen natürlichsprachlichen Zügen in künstlichen Indexsprachen. In M. Schönenberger, V. Engerer, P. Öhl, & B. Brogyanyi (Eds.), *Dialekte, Konzepte, Kontakte: Ergebnisse des Arbeitstreffens der Gesellschaft für Sprache und Sprachen, GeSuS e.V., 31. Mai – 1. Juni 2013 in Freiburg/Breisgau* (pp. 61-74). http://redaktion.gesus-info.de/S&S-online/S&S-Sonderheft_2014.pdf

Indexierungstheorie für Linguisten. Zu einigen natürlichsprachlichen Zügen in künstlichen Indexsprachen

Volkmar Engerer, Aalborg

Abstrakt
Die Informationswissenschaft hat eine lange Tradition von 'Sprachen' in der Indexierung von Dokumenten zu sprechen. Die sog. 'Indexierungssprachen' stellen dem Indexer einen Kode zur Verfügung, bestehend aus einem Lexikon von Indextermen und einer rudimentären Syntax, der es ihm ermöglicht, 'Themastatements' (inhaltliche Charakterisierungen von Dokumenten) zu formulieren, auf welche Benutzer durch Suchanfragen, vermittelt durch Information retrieval-Systeme, zugreifen können. Diese informationswissenschaftliche, aber dennoch sprachlich und kommunikativ geprägte Konstellation dient in vorliegendem Beitrag als Ausgangspunkt für eine Diskussion sprachlich-informationswissenschaftlicher Konzepte wie Syntax, Semantik oder paradigmatisch/syntagmatisch, und es wird danach gefragt, wie weit informationswissenschaftliche und linguistische Konzeptionen überlappen, divergieren oder füreinander nutzbar gemacht werden können. Der Begriff 'Indexsprache' wird erläutert und die Parallelität von Thesauri als kontrollierten Vokabularen und natürlichsprachlichen Lexika herausgearbeitet. Als konkretes Feld einer analogen Strukturierung werden Reihenfolgebeziehungen untersucht, welche in der Indextheorie als syntagmatische 'Zitierordnung' hervortreten, und in der Linguistik z.B. in der Abfolge der nominalen Attribute zum Vorschein kommen. Weiterhin werden die paradigmatisch-semantischen Beziehungen in Thesauri, welche grob in die drei Klassen Äquivalenzen, Hierarchien und assoziative Relationen fallen, den reichhaltigen semantischen Beziehungen in natürlichsprachlichen Lexika gegenübergestellt. Eine der hier vertretenen Schlussfolgerungen ist, dass die beiden Gebiete Informationswissenschaft und Linguistik sich viel zu sagen haben – und eine wirklich interdisziplinäre Annäherung erst in den Anfängen steckt.

Schlüsselwörter
Indexierungssprachen, Informationswissenschaft, künstliche Sprachen, Nominalgruppe, semantische Relationen, Thesauri

1 Indexsprachen

In vorliegendem Beitrag soll auf neuere Versuche, linguistische Konzepte in informationswissenschaftliche Forschung und Traditionen zu integrieren, eingegangen werden. Pandeys Untersuchung, die hier stellvertretend vorgestellt wird (Pandey 2003), ist ein Nachweis in Buchform, dass die Idee einer parallelen und dichotomen Strukturierung von Semantik und Syntax, welche in der Linguistik ein wichtiges Prinzip der Sprachbeschreibung darstellt, auch für das Design von *information-retrieval* (IR)-Systemen grundlegend ist. Wir haben es hier mit einer, aus der Sicht der Linguistik, traditionellen Fassung der Semantik-Syntax-Unterscheidung zu tun, die, wie übrigens auch schon beim

2014 Volkmar Engerer. Indexierungstheorie für Linguisten. Zu einigen natürlichsprachigen Zügen in künstlichen Indexsprachen. *Dialekte, Konzepte, Kontakte. Ergebnisse des Arbeitstreffens der GeSuS 2013 in Freiburg/Breisgau*, 61–74.
Kontakt: Volkmar Engerer, Universität Kopenhagen
e-mail: rhd237@iva.ku.dk

Informationswissenschaftler Frohmann (1990), ausschließlich in der Indexierungskomponente zum Tragen kommt – und nicht, wie in anderen Ansätzen, das kognitive Systems des Benutzers als semantischen "Verarbeitungsmechanismus" miteinschließt. In solchen systemorientierten und den bibliothekswissenschaftlichen Methoden in höherem Masse verpflichteten Ansätzen (Tedd 2005, Li 2009, Lancaster 2003) sind Semantik und Syntax den beiden grundlegenden Prozessen in der Inhaltsanalyse eines Dokuments parallelgeschaltet, nämlich 1. der Identifizierung von Konzepten, die in einem Dokument enthalten sind, und 2. der Bestimmung der Relationen, welche sie im Text verbinden (Pandey 2003: 23). Diese "doppelte Sprachanalyse", welcher schon die Idee einer paradigmatischen vs. syntagmatischen Achse innewohnt und die, wie noch zu zeigen ist, von J. Warner weitergedacht worden ist (Warner 2007a/b), kann grundsätzlich auf jegliches Sprachverstehen angewendet werden und stellt, in dieser Formulierung, keinen für den Indexierungsprozess spezifischen Begriff dar. Was traditionelle Indexierung von allgemeinem Textverstehen u.a. unterscheidet, ist nicht nur der Zweck (Indexierung zielt auf Suchbarkeit der Indexterme ab, wogegen Textverstehen keinem derartigen Ziel verpflichtet ist), sondern auch die inhaltstreue Reduktion von Textinhalt auf seine "Essenz", also eine verkürzte[1] Wiedergabe des Dokumentinhaltes durch Terme als Repräsentanten von Konzepten. Diese Konzepte/Terme, die entweder durch Extraktion aus dem Originaldokument oder durch Zuordnung aus einem externen Vokabular (Thesaurus) gewonnen werden, werden in einem zweiten Schritt durch syntaktische Relationen verbunden.[2] Die resultierende Kette aus syntaktisch verbundenen Termen stellt in Bezug auf die Dokumente, die sie repräsentiert, ein "Themastatement" dar (Pandey 2003: 30) und wird in der traditionellen Indextheorie als Repräsentation des Ausgangstextes angesehen. Indexierung umfasst aus dieser Sicht das Verstehen und die Repräsentation einer Textessenz (Fugmann 2002: 222). Themastatements sind so gesehen vergleichbar mit natürlichsprachlichen Sätzen und bestehen aus symbolischen Ketten, in denen Lexikoneinheiten (Wörter, Terme) in eine syntaktische Sequenz gebracht werden. In der Informationswissenschaft heißt

a) die Menge von Beschreibungstermen,
b) die Angabe einer relationalen Struktur in a) und
c) die Regelung der Kombinationsmöglichkeiten von Termen inklusive Reihenfolgebeziehungen in der Sequenz/im Themastatement

"Indexsprache". Die a)/b)-Komponente, die Terme und ihre Relationen, wird im Allgemeinen als "Thesaurus" bezeichnet (vgl. Broughton 2006, Foskett 1994), ein Begriff, der auch aus der

[1] Warner kritisiert diese Verkürzungstransformation mit Hinblick auf Indexierungstechniken im Bereich der Volltextsuche als veraltetes bibliothekswissenschaftliches Dogma, wenn er, in Anlehnung an Wilson (2001), feststellt: "The need for descriptions less extensive than the documents described, imposed by storage constraints of inscribed media, and for direct human intervention in the creation of these descriptions, [...] have tended to be universalized and treated as if they were independent of their dominant technological realizations [...]" (Warner 2010: 5f).

[2] Schon Sparck Jones & Kay stellten hierzu die rhetorische Frage an den Linguisten: "[I]s the kind of drastic compression of content that must be done to provide a document with an index description a process about which linguists can reasonably be expected to have anything to say?" (Sparck Jones & Kay 1973: 54)

Sprachwissenschaft bekannt ist (vgl. Lyons 1977: 300). Primitive Indexsprachen bestehen nur aus einer a)-Komponente, physisch existieren solche "Sprachen" dann als Wortlisten ("keyword lists") (vgl. Typ II1 in Broughton 2006: 20) - erst b) macht sie dann zu einem ausgewachsenen Thesaurus. Ein semantisch strukturiertes Vokabular im Sinne der informationswissenschaftlichen Indexierungstheorie (a/b) ist damit auch die Einheit, die einem natürlichsprachlichen Lexikon am nächsten kommt. Ein Thesaurus als Monolith von Lexikoneinheiten und einem Netzwerk aus a priori festgelegten Beziehungen zwischen ihnen kann alleine nicht ein Themastatement repräsentieren, eine Konkatenation innerhalb einer rudimentären Syntax des Verkettungstyps "+" ohne Reihenfolgeregelung ist das Minimum, um zu einer Themabeschreibung zu gelangen. Indexierungen, die mit einem Thesaurus und einer rein verkettenden Syntax ohne Reihenfolgebeziehungen arbeiten, erzeugen Statements der Form "A + B + ...", d.h. eine ungeordneten Termliste aus den Ausdrücken A, B, ... Sprachen mit dieser Art von verkettender Syntax sind in der Indexierungspraxis heutzutage mehr oder weniger der Standardfall, sind aber zu primitiv, um unter den natürlichen Sprachen vertreten zu sein.

Diese Form der Anordnung, die pure Auflistung von Beschreibungstermen im Themastatement, ist typisch für postkoordinierende Indexierung (Broughton 2006: 9, Weinberg 2009: 2283), wo erst der Benutzer, und nicht mehr der Indexer, durch die formale Gestaltung seines Suchausdrucks die syntaktischen Beziehungen zwischen den Suchtermen herstellt (einige semantische Konsequenzen dieser Benutzereinbeziehung sind von Warner 2007b aufgezeigt worden). Demgegenüber steht die präkoordinierende Methode, welche schon in der Indexierungsphase die syntaktische Gliederung, über den +-Typ hinaus, in einer voll ausgebauten c)-Komponente wahrnimmt. Letztere erst macht es möglich, die semantische Vokabularstrukturierung (a/b) direkt mit der syntaktischen Kombination in Themastatements (c) in Bezug auf ein bestimmtes Dokument zu verbinden. Erst unter den Bedingungen der Präkoordination wird das Ineinandergreifen von Semantik und Syntax mit allen seinen natürlichsprachlichen Konsequenzen und Parallelen direkt sichtbar, weshalb wohl Pandey seine Untersuchung auf diesen Typ begrenzt.

2 Kontrollierte Thesauri und natürlichsprachliche Lexika

Indexsprachen des präkoordinativen Typs operieren auf zwei Niveaus, 1. dem der Begriffe mit den dazugehörigen Termen und begrifflichen Relationen (a/b) und 2. dem der Kombinationen zwischen Termen einerseits und der begrifflich-kompositionellen Struktur, welche sich aus der Termkombination ergibt, andererseits (c). Diese Sprachauffassung kombiniert Lexikon-Thesaurus (a/b) und Syntax (c), was eine, auch in der theoretischen Linguistik, nicht unübliche Konzeption einer natürlichen Sprache ist.[3] Niveau 1, der

[3] Fraglich ist hier nur, ob der Begriff "Semantik" in seiner Anwendung auf den a/b-Komplex von einem strukturierten Lexikon nicht zu eng ist: Thesauri gehören in der Regel zu den "kontrollierten Vokabularen" (Lancaster 2003: 19), was heißt, dass der Thesaurus bestimmt, welche Wörter zur Sprache gehören und welche

Thesaurus, enthält alle semantischen kategorialen Relationen des paradigmatischen Typs, welche implizit als "Hintergrund" des Themastatements gegeben sind und sozusagen a priori existieren, bevor die Terme zur Beschreibung eines Dokumentinhaltes zusammengesetzt werden. Semantische Relationen (b), die in ein IR-Lexikon eingehen, werden allgemein in drei Gruppen (Äquivalenzen, Hierarchien und affinitive/assoziative Relationen) unterteilt und sind im Lexikon durch die vier Operatoren NT (*narrower term*), BT (*broader term*), RT (*related term*) und UF (*use for*) repräsentiert (Pandey 2003: 36, Broughton 2006: 23).

Aus der strukturellen Lexikontheorie sind diese Arten von Beziehungen gut bekannt (Lyons 1977: 270ff), und es fällt auf, dass die Parallelen zwischen einem "künstlich" designten, kontrolliertem Vokabular/Thesaurus (Lancaster 2003: 19) und der natürlichsprachlichen semantischen Strukturierung eines einzelsprachlichen Wortschatzes (Sparck Jones & Kay 1973: 169) nicht in ihrer vollen Tragweite gesehen wurden, obwohl eine Andeutung des Zusammenhangs schon bei Sparck Jones & Kay zu lesen ist:

> "Thus from the linguistic point of view the main feature of the typical term vocabulary is that it is ad hoc; it is oriented to a particular subject area, and individual term names represent specific or extended meanings of their ordinary language counterparts." (Sparck Jones & Kay 1973: 151)

Ein Grund für diese "Berührungsängste" könnte ein Dogma der Klassifikationstheorie sein, nämlich dass eine universal gültige, begriffliche Klassifikation sich an den Dingen in der Wirklichkeit zu orientieren hat (Batley 2005: 1–3), und nicht abhängig sein soll von den sprachlichen Bezeichnungen für diese Dinge. Aber Klassifikationen sind, wie schon Sparck Jones & Kay (1973: 158) angemerkt haben, oft nicht eindeutig in Bezug darauf, ob sie mit Wörtern, Wortbedeutungen oder außersprachlichen Konzepten hantieren, vgl.

> "[...] their relationship with vocabulary classifications, which are widely thought of as belonging to the lexicons of natural language grammars, is obscure." (Sparck Jones & Kay 1973: 159)

Mit diesen Bemerkungen sollen die Unterschiede zwischen einem Lexikon in einer natürlichen Sprache und einem Thesaurus im Sinne der Indexierungstheorie nicht unter den Tisch gekehrt werden, über die offensichtlichen und formalen Unterschiede hinaus (z.B. natürliches vs. kontrolliertes/konstruiertes Lexikon, Unterschiede in der inhaltlichen Breite sowie Diversität der semantischen Relationen zwischen Lexikoneinheiten usw.) ist es v.a. die Fixierung auf den Produktionsaspekt, der Thesauri von natürlichen Lexika unterscheidet. Während natürlichsprachliche Lexika das sprachliche Wortwissen eines Sprachteilhabers strukturieren, beschreiben und, wenn man will, lokal in einem mentalen, "internen" System repräsentieren, sind kontrollierte Thesauri externe, z.B. in Wortlisten fixierte Kontrollinstrumente für die Korrektheit von Themastatements und dienen v.a. dazu, den Indexer bei der Produktion von konsistenten, vollständigen usw. Dokumentbeschreibungen zu unterstützen und dem Benutzer bei der Auswahl der "richtigen" Suchterme behilflich zu sein,

---

Formen sie annehmen können. Der Begriff "Semantik" deckt diese formalen Aspekte wohl nicht im vollen Masse.

so dass seine Suche so effektiv und zufriedenstellend wie möglich verläuft (zum Begriff des "Suchthesaurus" vgl. Lancaster 1976: 30, Broughton 2006: 33). Diese beiden Funktionen bestimmen in hohem Masse das formale Design und die verwendeten Inhaltelemente eines Thesaurus – und setzen natürlichsprachliche Vokabulare rein funktionsmäßig, und trotz aller strukturellen Parallelen, von künstlichen Thesauri ab.

3 Reihenfolgebeziehungen in Themastatements vs. natürlichsprachliche Syntax

Niveau 2, die Syntax (syntagmatisch, syntaktische Relationen), enthält Regeln zur Zusammensetzung von komplexen Themen (Themastatements), welche in der Indexsprache selbst nicht enthalten sind und deshalb a posteriori erst im einzelnen Statement in Bezug auf ein spezifisches Dokument explizit gemacht werden (Pandey 2003: 35). Syntaktische Beziehungen in Indexsprachen sind von dreierlei Art, sie umfassen i) Reihenfolgebeziehungen, ii) eine Bestimmung der Art der involvierten Konzepte sowie iii) einen Mechanismus (Relatoren), der sie verbindet (Pandey 2003: 37).

Auch hier, im syntaktischen Bereich, sind Parallelen zwischen Indexsprachen und natürlichen Sprachen nicht voll ausgelotet worden. Konkret wird in der Indexierungstheorie die Syntax von Themastatements oft als "Zitationsordnung" verstanden, was die Reihenfolge semantischer Klassen von Termen in einem syntaktischen Ausdruck meint, z.B. an erster Position das spezifischste Konzept, und an den folgenden Positionen die ansteigend mehr generellen Konzepte (Batley 2005: 17). Klassifikationstheoretiker haben zu dieser formalen Ordnung eine Reihe von praktischen Vorschlägen gemacht, wovon Ranganathans 5 "Fundamentalkategorien" im Rahmen seiner "Facettenanalyse" (Broughton 2006: 10, Hjørland 2013) wohl die einflussreichsten sind. Ranganathans Fundamentalkategorien "Personality", "Matter", "Energy", "Space" und "Time", mnemotechnisch abgekürzt als PMEST (Pandey 2003: 53), zielen auf eine Ontologie von Themen ab, sind also zunächst als Klassen von Konzepten gedacht, auf welchen die Reihenfolge von Termen verschiedener Klassenzugehörigkeit in komplexen Themastatements in Übereinstimmung mit obiger Formel PMEST definiert ist (Pandey 2003: 58). Ein anderer Vorschlag stammt von Foskett (Foskett 1970, zit. nach Pandey 2003: 76–79), in dem die Zitationsordnung, das Problem der Sequenzierung von Konzepttermen, durch eine "Integrative Level Theory" geregelt ist, nach der die Reihenfolge der Konzeptterme in einem Themastatement parallel zu einer natürlichen Progression von Entitäten von niedrigeren zu höheren Niveaus der Organisation verläuft (Pandey 2003: 76). Diese Theorie ist später vom Prinzip der abnehmenden Generalität/zunehmenden Spezifizität ersetzt worden, das folgende Sequenzierung der Terme veranschlagt (Foskett 1970: 34f, zit. nach Pandey 2003: 78f):

> Relative Terme (z.B. Gradabstufungen) > Wertende Terme (positive/negative Einstellungen) > Positionen (Zeit, Raum, ...) > Physikalische Masse > Form > Erscheinung (Licht, Farbe) > Laute > Taktile Empfindungen > Geschmack > Gerüche > Zustände > Struktur

Vickerys "Standard-Zitationsordnung" (Pandey 2003: 95, vgl. auch Hjørland 2013) ist ein anderes Beispiel für Versuche, die Reihenfolge von Termen durch eine ontologische Gliederung zu motivieren.

In sprachlicher Hinsicht stößt man auf ähnliche syntaktische, positionsbezogene Argumente auf dem Gebiet der relativ freien Wortstellung von Adverbialen im Deutschen Mittelfeld (vgl. Macheiner 1991) oder auch bei der Frage der Sequenzierung von adjektivischen Attributen vor dem Kopfnomen im Deutschen (vgl. Eichinger 1982, 1991, sowie Posner 1980; zum Englischen vgl. Crystal 1976 und Dixon 1977). Untersuchungen zu diesen Themen sind deswegen interessant, weil sie die Grenzen zwischen grammatisch-reglementierten und bedeutungsbezogen-"freien" Stellungsregularitäten aufsuchen und auf einer semantischen Klassifizierung aufbauen, welche den Kriterien der Konzepttermabfolge in indexsprachlichen Ausdrücken nicht unähnlich ist. Obwohl letztere mit der ontologisch motivierten Reihenfolge der Konzeptterme in einer Indexsprache Bezug auf eine sprachunabhängige Gliederung der Welt nehmen (besser: wollen; das ist natürlich eine sprachphilosophisch äußerst heikle Annahme), wogegen linguistische Analysen auf semantische Eigenschaften der bezeichnenden Ausdrücke hinweisen, sind deutliche Analogien zwischen informationswissenschaftlichen ontologischen Bestrebungen und den semantischen Klassen sprachlicher Ausdrücke, hier Adjektiven, erkennbar, wie folgende Zusammenstellung von Fosketts positionsontologischen Klassen-Korrelationen (Foskett 1970, zit. nach Pandey 2003: 78f) sowie einer Synopse desselben Typs für Adjektivreihenfolgebeziehung in Eichinger (1991) zeigt:

Tabelle 1: Einige Parallelitäten in der Serialisierung von Indextermen in Themastatements (präkoordinierende Indexierung) und attributiven Adjektiven hin zum Kopfnomen.

	1	2	3	4	5
Indexsprachen (Foskett 1970)	Relative Terme (z.B. Grad-abstufungen)	Wertende Terme (positive/ negative Einstellungen)	Positionen (Zeit, Raum, ...)	Physikalische Masse, Form, Erscheinung (Licht, Farbe)	Laute, taktile Empfindungen, Geschmack, Gerüche, Zustände, Struktur
Natürliche Sprachen – deutsch (Eichinger 1991: 318f)	Quantifikatoren, Zahlausdrücke	Affektiv, evaluativ, Wertung, qualitativ	Zeitliche/räumliche Lage (Duden), vgl. aber Ort/Zeit (Sommerfeldt) sowie referierend (temporal/lokal) (Eichinger)[4]	Color (Seiler), Qualität (Sommerfeldt), qualitativ (Eichinger), Farbadjektive (Duden), qualifikative Adjektive (Engel)	Materiell (Seiler, fraglich), kategorisierend (konkret, materiell) (Eichinger), Stoff, Herkunft, Bereich (Duden), klassifikative Adjektive (Engel)

Eine Theorie, welche die Syntax von Indexsprachen mit der von natürlichen Sprachen wie hier im Bereich der konzeptuell/semantisch bedingten Stellungsregularitäten von Termen/Adjektivklassen in Beziehung setzen will, muss auch eine Reihe von offensichtlichen Unterschieden zwischen den beiden erklären. Hier seien fünf davon kurz angesprochen.

1. Die Reihung von Adjektiven in natürlichen Sprachen kann eine schrittweise semantische Anwendung nach dem Operator-Operand-Prinzip beinhalten, wie es z.B. mit dem äußersten Glied von Zahlausdrücken (vgl. Position 1, oben) der Fall ist; Ausdrücke dieser Art werden semantisch in einem Zug auf den gesamten Komplex der Eigenschaften (2–5) bezogen. Diese hierarchische Klammerstruktur und semantische Abarbeitung entlang einer positionellen Linie ist für die syntaktischen Abfolgen von Termen in Indexstatements nicht bekannt.
2. Die Kette von adjektivischen Attributen ist hin zum Kopfnomen strukturiert, reine Adjektivsequenzen ohne Kernsubstantiv sind in natürlichen Sprachen weder wohlgeformt noch kommunikativ verwendbar. Indexsprachliche Termsequenzen hingegen sind wohlgeformte Themastatements/Sätze ohne diese nominale Zielstruktur, also satzwertige Konstruktionen ohne Kopfelement.
3. Reihenfolgebeziehungen in natürlichen Sprachen sind immer auch von der untersuchten Einzelsprache abhängig, wogegen die Form von Themastatements einer Indexsprache weder von der natürlichen Spendersprache für das künstliche Lexikon, meistens Englisch, noch von der natürlich-einzelsprachlichen Provenienz der von Themastatements denotierten Dokumente abhängen sollte.

[4] Eichinger zitiert hier Sommerfeld, wo Ort/Zeit-Adjektive den wertenden (2) vorausgehen. Dieselbe Abweichung findet sich bei Eichingers eigenem Vorschlag.

4. Natürlichsprachliche Syntaxen sind Ergebnis wissenschaftlichen Beobachtens und entstammen einer deskriptiven Haltung zu einer Sprache, die, im Prinzip, auch ohne die Beschreibungen des Linguisten existiert (oder wieder besser: existieren sollte). Indexsprachliche Regeln dagegen entspringen bewussten normativen Akten von Informationsspezialisten, ohne welche die Sprache sowie ihre "Regeln" nicht existieren würden. Eine Theorie des Verhältnisses von Sprachen des einen und des anderen Typs hätte zu erklären, wie der Übergang von Normierung zu Beobachtung – und umgekehrt – zu interpretieren ist.
5. Die Anforderungen an syntaktische Ketten natürlichsprachlicher Ausdrücke und die syntaktische Gestaltung indexsprachlicher Statements sind verschieden. Es besteht eine kommunikative Disparität zwischen der, auf der einen Seite, natürlichsprachlichen Flexibilität von Sätzen, in ein ganzes Spektrum von Sprechhandlungen einzugehen, und der kommunikativen Spezialisierung von Themastatements auf der anderen. Natürlichsprachliche Ketten von Adjektiven sind immer Teil einer größeren syntaktischen Struktur, nämlich von Nominalgruppen (vgl. auch Einwand 2), die ihrerseits Teile von Sätzen sind und erst qua dieser Struktur in minimale kommunikative Handlungen eingehen können, welche z.B. von der Sprechakttheorie beschrieben werden (Austin 1989, Cole & Morgan 1975, Huang 2006, Searle 1975/1985). Indexsprachliche syntaktische Ausdrücke dagegen sind direkt kommunikativ in dem Sinne, dass ein Themastatement mit der gewählten Reihenfolge der Terme für einen Benutzer hilfreich sei, indem es erstens, ganz in Analogie zur repräsentierenden Funktion in assertiven Sprechakten, inhaltlich zutreffend ist (thematische Adäquatheit in Bezug auf die denotierten Dokumente), und es zweitens dem Benutzer, z.B. durch Browsen (Batley 2005: 137ff), erlaubt, schrittweise zu einer Kette zu gelangen, welche die für ihn relevanten Dokumente denotiert, diese Repräsentation also in einer Form kommuniziert, welche die Findbarkeit von Dokumenten in der Datenbank sichert.

Wie ausgeprägt der Parallelismus zwischen Indexierungssprachen und natürlichen Sprachen in Bezug auf die formale Ebenenorganisierung von Semantik und Syntax und die inhaltliche Charakterisierung dieser Ebenen dennoch ist, und in welch hohem Masse IR-Werkzeuge nach linguistischen Prinzipien organisiert sind, soll am Beispiel der Indexsprache PRECIS (*Preserved Context Indexing System*) demonstriert werden.[5]

[5] Vgl. hierzu Pandey (2003: 115–124), der sich auf das Handbuch zu PRECIS (Austin & Dykstra 1984) stützt. Allgemeinere Diskussionen der formalen Struktur von Indexsprachen mit Hilfe von linguistischen Kriterien sind Noel (1972) und Hutchins (1975), die in dieser Besprechung nicht mehr berücksichtigt werden können.

4 Semantische Relationen in einer Indexsprache (PRECIS) sowie in natürlicher Sprache

In der PRECIS-Syntax wird das in Frage stehende Thema erst in verschiedene Komponenten aufgespalten ("Facetten"), die dann systematisch nach einem Muster und generellen Regeln der Reihenfolgebeziehungen wieder zusammengesetzt werden. Die Klassen von Konzepten, die in das Vokabular der Sprache eingehen, sind im "key system" organisiert und enthalten physische, zählbare und nicht-zählbare Objekte (Telefone, Granit), Abstrakta (Hitze, Zeit), Systeme abstrakter Entitäten (Christentum) und Körperorgane (Herz) (Pandey 2003: 118). Das linguistische Pendant zu dieser Klassifikation sind semantische Wortartdefinitionen (hier von Substantiven) bzw. Versuche, grammatisch relevante Wortklassen auch semantisch zu motivieren. Bekannt ist die in natürlichen Sprachen relevante zählbar/nicht-zählbar-Unterscheidung, die Distinktion zwischen Konkreta und Abstrakta, ebenso wie "Systeme abstrakter Entitäten", die in sprachlicher Hinsicht durch Kollektive bzw. Institutionen denotierende Substantive (*Polizei*) vertreten sind. Auch die angeführte Klasse der Körperorgane hat sprachliche Reflexe, da Körperorgane in Bezug auf die Person in eine spezielle Teil-Ganzes-Relation eingehen, und, u.a., zur Gruppe der nicht-alienablen (nicht-veräußerbaren) Possessionsverhältnisse gehören (Lyons 1977: 311ff) und oft grammatikalisiert sind. Darüber hinaus werden in den Kernkonzepten ("core concepts") von PRECIS transitive von intransitiven Handlungen ("actions") sowie verschiedene Typen Ausübender einer "transitiven Handlung" (Agens, Instrument) unterschieden, alles Distinktionen, welche in der Grammatik natürlicher Sprachen, z.B. in Form von transitiver/intransitiver Verben oder semantischer Rollen wiederkehren.

Die Semantik von PRECIS ist in einem Thesaurus formuliert, der Terme mithilfe von drei grundlegenden lexikalischen Relationen verbindet (Pandey 2003: 121) (die drei Typen semantischer Relationen wurden oben schon kurz angesprochen):

1. **Äquivalenzen** verbinden Terme, die auf dasselbe Konzept referieren, auch "Synonyme" genannt. Eine Differenzierung geschieht auf den Dimensionen präferiert/nicht-präferiert, populär vs. wissenschaftlich, modern vs. veraltet usw.
2. **Hierarchien** (Pandey 2003: 122) verbinden Terme durch Über-/Unterordnung. Hier werden drei Subklassen unterschieden: a) Genus-Spezies ('Dokumente'-'Bücher'), b) Teil-Ganzes-Relationen ('Dänemark'-'Skandinavien'-'Nordeuropa'-...) und c) die Instantiierungsrelation, welche Klassen von Dingen, bezeichnet durch einen Gattungsnamen, mit einem Exemplar dieser Klasse verbindet, üblicherweise bezeichnet durch einen Eigennamen ('Linguist'-'Peter Öhl').
3. **Assoziative Relationen** (Pandey 2003: 123) stellen eine Verbindung her zwischen zwei Termen, die a) nicht zu einer Äquivalenzmenge gehören und b) auch nicht in einer Hierarchie verknüpft werden können, so dass der eine Term dem anderen untergeordnet ist. Stattdessen sind zwei Terme A und B "mental" miteinander assoziiert/verbunden, wenn angenommen werden kann, dass, nun dezidiert aus der Sicht der Informationswissenschaft argumentiert, ein Benutzer, der mit dem Term A

sucht, mit gewisser Wahrscheinlichkeit auch mit einer Suche anhand des Terms B gedient sein kann. Beispiele für assoziative Beziehungen sind die Relationen Disziplin-Studienobjekt der Disziplin, genetische Verhältnisse, hierunter auch sprachliche ('Indogermanisch'-'Deutsch'), Behälter-Inhalt ('Kanne'-'Milch'), Handlung-Resultat ('Weben'-'Stoff'), Ursache-Wirkung, Prozess-Agens/Instrument ('Schreiben'-'Schreibmaschine'), Handlung-Patiens/Objekt ('Ernte'-'Korn'), ...

Ich komme nun auf einige sprachlich-grammatische Gegenstücke dieser, ursprünglich für den Indexer formulierten, semantischen Relationen zu sprechen. Die Gruppe von Äquivalenzen (vgl. 1) kann verbunden werden mit dem traditionellen Gebiet der linguistischen Synonymforschung, der funktionellen Differenzierung semantisch ähnlicher Lexikoneinheiten nach Stil, Stratum, fachsprachlichen Eigenentwicklungen oder sozialen Merkmalen der Sprechergruppe. Hierarchien (vgl. 2) bilden, wie schon mehrmals angedeutet, das Rückgrat der semantischen Strukturierung eines natürlichsprachlichen Wortschatzes, in erster Linie als Hyponymie-Beziehung zwischen Wortschatzeinheiten, welche die Genus-Spezies-Relation (2a) in Semantiken von Indexsprachen/Thesauri sprachlich realisiert. Diese Beziehung (vgl. für das Folgende Lyons 1977: 291–293) kommt z.B. in einer einseitigen Implikationsbeziehung zwischen einem Satz mit Hyponym und dem entsprechenden Satz mit dem übergeordneten Lexem zum Ausdruck (z.B. "Else ging mit ihrem Labrador spazieren" -> "Else ging mit ihrem Hund spazieren"), oder, um einen anderen sprachlichen Reflex zu nennen, in Satzmustern des Typs "x ist eine Art (von) y", wo die Einsetzung von Lexemen für x und y folgender Regel von Über- und Unterordnung genügen muss: <x: Hyponym, y: übergeordnetes Lexem>, z.B. "Ein Labrador ist eine Art (von) Hund". Ebenso wie die hierarchische Genus-Spezies-Relationen in Thesauri ist die sprachliche Hyponymie-Relation transitiv: Wenn x Hyponym ist zu y, und y Hyponym zu z, dann ist auch x Hyponym zu z. Unterklassen eines Genus werden aus der Sicht der Thesaurustheorie zu Spezies, indem zum Genusmerkmal eine differenzierende Eigenschaft hinzugefügt wird. In natürlichen hierarchischen Wortschatzbeziehungen ist in gleichem Zusammenhang von "Ko-hyponymen" (mehrere Hyponyme unter einem übergeordneten Begriff) die Rede, die in sprachlicher Hinsicht als syntaktische Ausdrücke aus einem attributivem Modifikator und dem übergeordneten Lexem analysiert werden können ('Tyrann' = 'despotischer Herrscher'); dies ist auch die Standardform für Definitionen.

Auch die Teil-Ganzes-Relation (2b) ist in ihrer Relevanz für Lexikonstrukturen sowie syntaktische Konstruktionstypen wie die Possession in natürlichen Sprachen belegt (Lyons 1977: 311–315). Die Instantiierungsrelation (2c) scheint in natürlichen Sprachen vorzugsweise nicht lexikalisiert zu sein, im Deutschen wird das Verhältnis von Individuum und Klasse (bzw. Klassenzugehörigkeit) durch den Konstruktionstyp des *ist*-Prädikatsatzes ("Peter Öhl ist Linguist") oder auch appositive Konstruktionen wie "der Linguist Peter Öhl"/"Peter Öhl, der Linguist" ausgedrückt.

Assoziative Beziehungen (vgl. 3 oben) hingegen werden im Deutschen oft mit morphologischen Mitteln identifiziert – was sie somit mit einem Bein in das Lexikon stellt.

Die assoziative Beziehung zwischen einem Studiengegenstand und der entsprechenden Disziplin ist oft vermittelt durch ein Paar <Simplex/Gegenstand, Komposition (Erstglied: Gegenstand, Zweitglied: -wissenschaft)> wie in <Sprache, Sprachwissenschaft>, <Natur, Naturwissenschaft>, usw. Die assoziative Relation von <Handlung, Resultat> kann im Deutschen einen Ablautreflex haben (z.B. <Schreiben, Schrift>). Die Relation <Handlung, Patiens> drückt sich in einer Reihe von Beispielen in Verbalsubstantivierungen mit entsprechenden substantivierten Partizip Passivformen als Zweitpartner aus wie in <Verhaftung, Verhafteter> oder <Bestrafung, Bestrafter>. Ein weiteres Beispiel ist das assoziative Paar von <Substanz, Eigenschaft>, welche, auch wieder im Deutschen, durch *-ig/ -lich*-Ableitungen bewerkstelligt werden kann (man bemerke, dass sprachliche Realisierungen von paradigmatischen Relationen nicht, wie in informationswissenschaftlichen Thesauri, auf Wortartidentität zwischen den, in der Regel, substantivischen Relationspartnern beschränkt sind), z.B. <Gift, giftig>, <Stoff, stofflich>, ...

Die Gliederung in eine syntaktische und eine semantische Ebene der Indexierung ermöglicht auch in traditionellen Auffassungen der Indextheorie die Parallelisierung grundlegender Strukturierungsprinzipien in Index- und natürlichen Sprachen – unter gleichzeitiger Respektierung basaler Unterschiede, welche sowohl die Existenzweise (künstlicher, geschaffener Charakter vs. Natürlichkeit, mentale Verankerung) als auch die Funktion (Repräsentationsfunktion und kommunikatives "Angebot" an den Benutzer vs. Vielzahl von kommunikativen Aufgaben) von Indexsprachen sowie natürlichen Sprachen betreffen. Pandey, dem es vorderhand um eine Konsolidierung der Rolle seines indischen Landsmannes Ranganathan in der Etablierung linguistischer Prinzipien in der Klassifikationstheorie geht, hält Ranganathans Theorie einer Indexsprache für gültig auch in automatisierten IR-Systemen, auf jeden Fall, was die Semantik angeht. Die syntaktischen Prinzipien von Ranganathans Theorie betrachtet Pandey (2003: 131) allerdings nicht als anwendbar auf post-koordinierende Indexsprachen. Mit Überlegungen dazu schliesst mein Beitrag ab.

5 Perspektiven: die zukünftige Rolle sprachwissenschaftlicher Konzepte in der Indexierungstheorie

Es bleibt also die abschließende Frage, ob die Rolle der Linguistik mit den dominierenden post-koordinativen Indexierungssystemen, die alle syntaktischen Tätigkeiten dem Benutzer auferlegen, und mit dem Einzug von Volltext-Techniken, welche die Notwendigkeit einer, wie immer gearteten verkürzten Repräsentation und Indexierung in ihrer Gesamtheit in Frage stellen, ausgespielt ist. Ein radikalerer Versuch, linguistische Denkweisen auf technologische Up-to-date-IR-Systeme mit post-koordinativen Indexsystemen auf Volltextbasis anzuwenden, stammt von J. Warner, die er in seinem Buch "Human information retrieval" (Warner 2010) vorgelegt hat. In zwei vorbereitenden Arbeiten (Warner 2007a/b) koppelt er die Zweiteilung einer syntaktischen und semantischen Ebene, die bei Pandey noch auf formale Indexsprachen beschränkt war, an die Idee einer paradigmatischen vs. syntagmatischen Achse in IR-

Systemen (vgl. Saussure 1967, Lyons 1977: 270), und bezieht den Benutzer, bzw. sein kognitives System, als "linguistischen" Akteur explizit mit ein. Als Fallbeispiel dient Warner die automatische Indexierung (Sparck Jones & Kay 1973: 10, 29, 63, Mai 1999: 276, 287, Weinberg 2009: 2286, Tedd 2005: 170), in welcher ein Algorithmus aus einem Volltext-Dokument suchbare Beschreibungsterme extrahiert. In Saussures Dichotomie von Syntagmatik/Paradigmatik kann dies als das Herauslösen eines Elements, hier eines Wortes, aus seinem syntagmatischen Kontext in einem spezifischen Dokument verstanden werden (Warner 2007b: 275). Solche zu Indextermen transformierten Wörter sind im Index sozusagen semantisch "freigesetzt" und in ihr größtmögliches Paradigma "losgelassen", sie werden maximal mehrdeutig. Erst in Suchabfragen durch Benutzer werden diese Einheiten in einem Abfragestatement neu kombiniert – was einer "Wiedereinsetzung" vordem isolierter Terme (mit einer Vielzahl von paradigmatischen Bedeutungen) in ein neues Syntagma entspricht, welches dann in den gefundenen Dokumenten und Textstellen in eine Vielzahl syntagmatischer Vorkommnisse "rückübersetzt" wird. Mit den Worten von Warner handelt es sich hier um eine umgekehrte Transformation eines Paradigmas (Warner 2007b: 275). Es ist in dieser Umkehrung möglich, dass die Bedeutungen von Suchtermen, welche erst in dem vollständigeren Syntagma des jeweiligen Dokuments hervortreten, nicht mit den intendierten Bedeutungen des Benutzers übereinstimmen (Warner 2007b: 275f). Erst durch das "Retrieval" von Textdokumenten werden also die syntagmatischen Umgebungen von Abfragetermen wiederhergestellt, und die Spannweite der paradigmatischen Bedeutungen eines einzigen Abfrageterms wird durch variierende syntagmatische Umgebungen abgesteckt (Warner 2007b: 276):

> "Linguistics can, then, contribute a sophisticated understanding of the interaction between signifier and signified enforced by the movement in description from syntagm to paradigm, and from paradigm to syntagm in searching and retrieval, for computational and direct human operations on written language in full-text representation and retrieval." (Warner 2007b: 276)

Es handelt sich hier ohne Zweifel um einen innovativen Ansatz, der nicht nur strukturelles linguistisches Gedankengut für eine neue Sicht auf aktuelle informationswissenschaftliche Problemstellungen nützlich macht, sondern auch, manchmal auch in provozierender Weise, traditionelle Sichtweisen auf das IR-Paradigma in Frage stellt, so z.B. das Dogma der "query transformation" (Warner 2010: 3). Während die mehr konservativen Versuche wie Pandey sprachliche und linguistisch motivierte Strukturierungen sozusagen "wiederentdecken", indem sie die sprachlichen Prinzipien herausarbeiten, die traditionellen IR-Theorien zugrunde liegen, "entdeckt" Warner linguistische Konzepte und sprachliche Analogien und erschließt damit neue Zugänge zu bekannten Problemen in der IR-Forschung. Es muss noch abgewartet werden, inwieweit diese Herausforderungen von der Mainstream-Informationswissenschaft aufgenommen werden.

Literatur

Austin, Derek & Mary Dykstra. 1984. *PRECIS: a manual of concept analysis and subject indexing.* 2. ed. London: British Library. Bibliographic Services Division.

Austin, John. L. 1989. *How to do things with words.* 2. ed., reprint. Oxford: Oxford University Press.

Batley, Susan. 2005. *Classification in theory and practice.* Oxford: Chandos Publishing.

Broughton, Vanda. 2006. *Essential thesaurus construction.* London: Facet.

Cole, Peter & Morgan, Jerry L. 1975. *Syntax and semantics: speech acts.* New York: Academic Press.

Crystal, David. 1976. *Linguistics.* Harmondsworth: Penguin.

Dixon, Robert M.W. 1977. Where have all the adjectives gone? *Studies in Language* 1.1,19–80.

Eichinger, Ludwig M. 1991. Ganz natürlich - aber im Rahmen bleiben. Zur Reihenfolge gestufter Adjektivattribute. *Deutsche Sprache* 19, 312–329.

Eichinger, Ludwig M. 1982. *Syntaktische Transposition und semantische Derivation: die Adjektive auf* –isch *im heutigen Deutsch.* Tübingen: Max Niemeyer Verlag.

Foskett, Douglas John. 1994. Thesaurus. In K. Sparck Jones & P. Willet (eds.), *Readings in Information Retrieval*, 111–134. San Francisco: Morgan Kaufmann.

Foskett, Douglas John. 1970. *Classification for a general index language: A review of recent research by the Classification Research Group.* (London): The Library Association.

Frohmann, Bernd. 1990. Rules of indexing: a critique of mentalism in information retrieval theory. *Journal of Documentation* 46. 2, 81–101.

Fugmann, Robert. 2002. The complementarity of natural and index language in the field of information supply: an overview of their specific capabilities and limitations. *Knowledge Organization* 29. 3/4, 217–230.

Hjørland, Birger. 2013. Facet analysis: The logical approach to knowledge organization. *Information Processing & Management* 49, 545–557.

Huang, Y. 2006. Speech Acts. In Keith Brown (ed.), *Encyclopedia of Language & Linguistics* (2nd ed.), 656–665. Oxford: Elsevier.

Hutchins, W. J. 1975. *Languages of indexing and classification - A linguistic study of structures and functions.* Stevenage: Peter Peregrinus Ltd.

Lancaster, F. Wilfrid. 2003. *Indexing and abstracting in theory and practice.* 3rd. ed. London: Facet.

Lancaster, F. Wilfrid. 1976. The relevance of linguistics to information science: perspective paper prepared for the FID. Oslo: Norsk senter for informatikk.

Li, LiLi. 2009. *Emerging technologies for academic libraries in the digital age.* Oxford: Chandos.

Lyons, John. 1977. Semantics. London: Cambridge University Press.

Macheiner, Judith. 1991. *Das grammatische Varieté oder die Kunst und das Vergnügen, deutsche Sätze zu bilden.* Frankfurt am Main: Eichborn.

Mai, Jens-Erik. 1999. Deconstructing the Indexing Process. *Advances in Librarianship* 23, 269–298.

Noel, Jacques. 1972. *A Semantic Analysis of Abstracts Around an Experiment in Mechanized Indexing.* Ph.D. Dissertation, Université de Liège.

Pandey, R. C. 2003. *Information retrieval system. A linguistic study.* Delhi: Abhijeet Publications.

Posner, Roland. 1980. Ikonismus in der Syntax. Zur natürlichen Stellung der Attribute. *Zeitschrift für Semiotik* 2.3, 57–83.

Saussure, Ferdinand de. 1967. *Grundfragen der allgemeinen Sprachwissenschaft.* Berlin: Walter de Gruyter.

Searle, John R. 1985. *Speech acts: an essay in philosophy of language.* Reprint. Cambridge.

Searle, John R. 1975. A taxonomy of illocutionary acts. In Keith Gunderson (ed.), *Language, mind and knowledge*, 344–369. Minneapolis: University of Minnesota Press.

Sparck Jones, Karen & Kay, Martin. 1973. *Linguistics and information science.* New York, London: Academic Press.

Tedd, Lucy A. 2005. *Digital Libraries: Principles and Practice in a Global Environment.* Berlin: Saur.

Warner, Julian. 2010. *Human information retrieval.* Cambridge, Mass.: MIT Press.

Warner, Julian. 2007a. Analogies between linguistics and information theory. *Journal of the American Society for Information Science and Technology* 58. 3, 309–321.

Warner, Julian. 2007b. Linguistics and information theory: Analytic advantages. *Journal of the American Society for Information Science and Technology* 58. 2, 275–285.

Weinberg, Bella Hass. 2009. Indexing: History and Theory. In Marcia J. Bates & Mary Niles Maack (eds.) *Encyclopedia of Library and Information Sciences.* 3rd. ed.: 2277-2290. New York: Marcel Dekker.
Wilson, Patrick. 2001. The Intellectual Foundations of Information Organization (Book Review). *College & Research Libraries* 62. 2., 203-204.